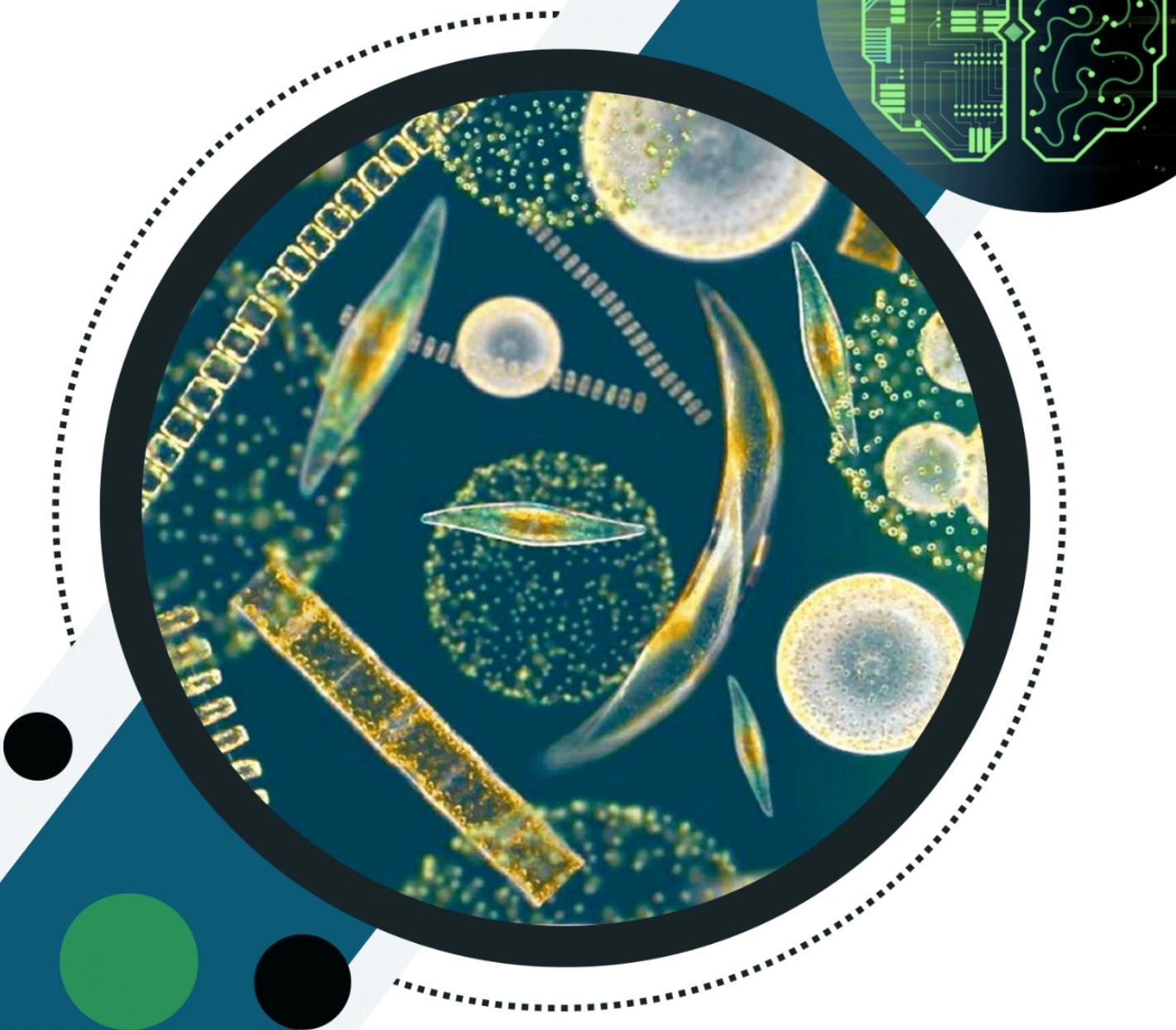


MACHINE LEARNING POUR LA COMPRÉHENSION DU PHYTOPLANCTON

Rapport Final



Réalisé par :

LITOU Alice / SMOCH Théo / SAADI Rayane
Becard Robin / GOETINCK Alexandre

Sommaire

1-Introduction.....	3
2-Machine Learning.....	4
3-Base de données.....	5
3.1-Présentation de la base de donnée.....	5
3.2-Prétraitement des données.....	5
4-Méthodes de classification.....	6
4.1-ACP	6
4.2-K-Means.....	9
5-Interprétation des résultats.....	13
6-Apports du projet.....	14
6.1-Apports personnels.....	14
6.2-Apports techniques.....	14
7-Conclusion.....	14
Remerciements.....	15
Annexes.....	16

1- Introduction

La biologie marine est au cœur de tout, aussi bien pour la régulation de la température du globe, du CO₂ et la biodiversité.

En effet, la biologie marine voit par centaines de milliers, voire des millions, le nombre d'interactions qui la constituent. C'est au centre de ces interactions que se placent les « phytoplanctons ».

Organismes multicellulaires végétaux, ils sont à la base de la chaîne alimentaire aquatique. Ils sont répartis en sept groupes, comme les diatomées, les micro algues, les cyanobactéries, les dinoflagellés...

Majoritairement constitués de Chlorophylle-a (permettant la photosynthèse) ainsi que de plusieurs autres pigments différents d'un groupe à un autre (par exemple la Chlorophylle-a total, Divinyl Chlorophylle-a, Chlorophylle-b, Péridinine ...). La présence ou non d'un type de pigment a des conséquences notables sur son environnement. L'un des facteurs les plus visibles sur lequel agit le phytoplancton et sa composition se trouve être la couleur de l'océan, qui varie ainsi du bleu profond des eaux pures au vert des eaux riches en phytoplancton ou en sédiments. Nous pouvons alors utiliser certaines de ces informations en tant qu'indicateurs, ce qui est le cas de la Chlorophylle-a qui est habituellement utilisée comme indicateur de la quantité de biomasse.

Ainsi, la classification des phytoplanctons est possible en fonction des pigments qui les composent. C'est dans ce cadre que se tient notre projet multidisciplinaire, son but: regrouper et classer les types de phytoplancton et de pigments afin de remarquer leurs différentes interactions. Pour cela il existe un outil qui simplifiera énormément la tâche de l'homme, il s'agit de l'intelligence artificielle et plus particulièrement le «Machine Learning», permettant à un programme d'apprendre par lui-même à réaliser notre tâche. Pour ce projet nous nous concentrerons donc sur les algorithmes « K-Means » ainsi que «ACP».

2- Machine Learning

Le Machine Learning est une catégorie de l'intelligence artificielle, qui a pour principal intérêt de "laisser" la machine apprendre par elle-même. L'intérêt est de ne pas devoir programmer l'intégralité de son comportement, mais donc de la laisser le développer, en utilisant divers algorithmes capables d'extraire et trier les données pour qu'elles puissent être exploitées et intégrées par l'IA.

Intéressons-nous tout d'abord au fonctionnement basique du Machine Learning. Nous pouvons relever 3 principaux "types" de Machine Learning : supervisé, non supervisé, et renforcement.

Le Machine Learning supervisé trie les données qu'on lui procure en selon leur appartenance à un groupe déjà présent dans une base de données. À l'inverse, le non supervisé trie également les données par classe, mais il n'est pas en possession d'une base de données initiales, il va tout regrouper et trier de lui-même. Enfin, le Machine Learning par renforcement est plutôt différent des deux autres: en effet, lui va effectuer des actions et en fonction de l'utilité de ce qu'il a fait, il recevra soit une punition, soit une récompense, et reproduira ses essais en tentant de maximiser les récompenses et minimiser les punitions.

C'est donc entre ces trois types de Machine Learning que nous devons choisir pour mener à bien notre projet. Étant en possession d'une base de données composée de données non triées sur le phytoplancton, qui sont des relevés pris dans différentes stations du globe. Il apparaît alors que la méthode la plus appropriée est le Machine Learning non supervisé. Mais alors, comment s'y prendre exactement ?

Nous avons abordé plus tôt la notion d'algorithmes; c'est en utilisant ces derniers que les trois types de Machine Learning fonctionnent. Dans notre cas, il faut utiliser ceux adaptés à du non supervisé. Il en existe un certain nombre, comme le Dimensionality Reduction, le Singular Value Decomposition, l'Indépendant Component Analysis... Il en existe beaucoup, mais les deux qui nous intéressent le plus, et qui sont les plus adaptés à notre cas précis sont le K-Means, et le Principal Component Analysis (PCA), ou Analyse des Principaux Composants (ACP).

Pour la bonne réalisation de ce projet, nous avons utilisé ces deux algorithmes parce qu'ils sont en réalité complémentaires.

3- Base de données

3.1-Présentation de la base de données

Pour effectuer la classification du phytoplancton et de leurs pigments nous disposons de deux bases de données, la première étant la base de données globale qui répertorie les concentrations des différents pigments prélevés à un endroit précis, la seconde étant la base de données de Tara Oceans qui comporte la concentration en pigments mais aussi l'abondance des différents types de phytoplanctons présents.

Dans le cadre de ce projet, nous nous sommes principalement intéressés à la première base de données, afin de travailler efficacement sur celle-ci et de trouver les bons regroupements pour nos différents pigments. Cette base de données est à l'origine composée de 9 484 lignes correspondant à autant de prélèvements effectués tout autour du globe (annexe 2). Pour chaque prélèvement, nous connaissons sa latitude, sa longitude, le jour, le mois et l'année du prélèvement, ainsi que la concentration (en mg/m^3) de chaque pigment le composant.

Cela représente une grande quantité de données, mais en l'état la base de données n'est pas utilisable à cause de plusieurs facteurs comme des valeurs manquantes, des valeurs aberrantes, ou encore une quantité importante de valeurs nulles qui empêchent la bonne analyse de cette base de données. Il faut donc avant tout la préparer pour l'analyse. Cela a donc été notre premier objectif lors de ce projet.

3.2-Prétraitement des données

Dans le but d'optimiser notre travail et d'identifier les regroupements appropriés pour les différents pigments étudiés, il est nécessaire de procéder à certaines modifications préliminaires sur cette base de données. Plus précisément, nous avons entrepris les étapes suivantes pour préparer les données de manière adéquate. Tout d'abord, nous avons éliminé les données manquantes et les colonnes non pertinentes, notamment celles relatives à la latitude, la longitude et la date de prélèvement. Cette étape de nettoyage nous a permis de nous concentrer uniquement sur les informations essentielles pour notre analyse, soit ici les différentes concentrations de pigments.

Ensuite, nous avons effectué une étude approfondie de la base de données afin de détecter et d'exclure les valeurs aberrantes. Cette démarche vise à garantir la qualité et la fiabilité des données utilisées dans notre traitement. Pour ce faire, nous avons déterminé

quelles étaient les concentrations supérieures à 95% des valeurs et celles inférieures à 5% et nous les avons exclues de notre base de données.

Étant donné que nous travaillons avec des échantillons de données relativement petits, nous avons pris la décision de les amplifier en utilisant le log10. Cette transformation rend les données plus lisibles et facilite leur interprétation. Cependant, une attention particulière doit être portée aux valeurs qui tendent vers moins l'infini après la transformation logarithmique. Afin de maintenir la cohérence des données, nous avons préféré exclure ces données plutôt que de les remplacer par nos plus petites valeurs, cette dernière méthode nous créait d'autres valeurs aberrantes et faussait notre analyse. Cette approche garantit donc que les données amplifiées restent dans des plages raisonnables et évite toute distorsion dans notre analyse.

Cette partie de prétraitement de la base de données a été une des parties qui nous a pris le plus de temps, car une seule erreur peut fausser toute la suite, il faut donc être très rigoureux et vérifier à chaque fois la cohérence de nos valeurs grâce à des graphiques que vous pourrez consulter en annexe, notamment ceux qui illustrent l'avant et l'après prétraitement en annexe 3,4.

4- Méthodes de classification

Lorsqu'on utilise le Machine Learning de manière non supervisée, nous sommes amenés à classer les individus par catégories afin de rendre le traitement par Machine Learning plus efficace. Il existe donc différentes méthodes de classification par algorithme, les méthodes que nous utiliserons ici sont (comme dit précédemment) les K-Means et l'algorithme ACP.

4.1-La méthode ACP (*Analyse en Composantes Principales*)

L'algorithme ACP cherche tout d'abord à représenter dans un graphique (avec un système de coordonnées X-Y) les valeurs de la matrice initiale. Il va ensuite tracer deux axes, le premier passant le long des valeurs présentant la plus grande variation, et un deuxième axe ayant la deuxième direction la plus importante, et il est orthogonal au premier axe.

Cet algorithme permet de se faire une idée de la redondance des données, en fonction du graphique qu'on obtient, et d'identifier les variables similaires.

Ainsi, après le traitement initial de la base de données pour supprimer les données non pertinentes, nous avons utilisé l'algorithme ACP. Il a fallu dans un premier temps supprimer , du tableau, des colonnes qui n'étaient pas utiles, comme la date ou les coordonnées géographiques. Comme il ne restait que les données réellement utiles, il était maintenant possible d'utiliser l'ACP.

Comme expliqué plus tôt, cet algorithme a pour fonction la réduction de dimension de bases de données, ce qui va ici nous permettre d'identifier les éléments les plus importants pour l'exploitation des résultats et l'utilisation de l'algorithme des K-Means.

Donc, après utilisation de l'algorithme, nous obtenons le tableau ci-dessous :

	Dimension	Variance expliquée	% variance expliquée	% cum. var. expliquée
0	Dim1	5.131035	57.0	57.0
1	Dim2	1.498030	17.0	74.0
2	Dim3	0.850624	9.0	83.0
3	Dim4	0.502289	6.0	89.0
4	Dim5	0.358758	4.0	93.0
5	Dim6	0.303413	3.0	96.0
6	Dim7	0.236632	3.0	99.0
7	Dim8	0.087547	1.0	100.0
8	Dim9	0.044812	0.0	100.0

Figure 1 - Tableau après l'utilisation de la méthode ACP

Ces 9 dimensions créées sont des dimensions arbitraires calculées à partir de la variance entre les différents pigments. L'élément le plus important de ce tableau est le pourcentage de variance expliquée: il nous permet de nous représenter l'importance des dimensions. On remarque assez vite que la première dimension est de loin la plus importante, avec sa variance expliquée s'élevant à 57%, suivie de la deuxième dimension avec une variance de 17%, et les autres étant en dessous de 10%.

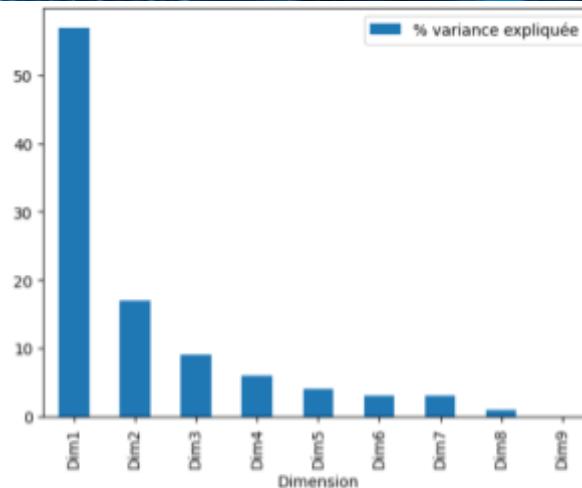


Figure 2 - Tableau des pourcentages de variance expliquée

Le tableau précédent nous permet de mieux nous représenter l'importance des dimensions. Nous pouvons ainsi, à partir de cette analyse, déterminer sur quelles dimensions il est plus intéressant de travailler, ici nous avons choisi d'en prendre 3, afin de travailler sur plusieurs dimensions tout en n'en traitant pas d'éléments peu utiles.

Ainsi, en ne gardant que les valeurs des 3 dimensions qui nous intéressent on obtient un tableau ressemblant à ceci :

	Dim1	Dim2	Dim3
0	3.912831	2.685004	-0.372297
1	3.952900	2.283444	-0.774588
2	4.383539	1.874272	-0.489254
3	2.933331	1.573114	-0.204273
4	2.743086	1.476750	-0.297398
...
681	2.753948	1.480291	-0.432457
682	2.799314	1.493896	-1.180147
683	3.258882	1.292053	-0.621750
684	3.119590	1.453788	-0.677135
685	5.153457	2.216067	-0.630034

Figure 3 - Réduction à 3 dimensions

Ce qui nous permet d'obtenir des nuages de points, que l'on réalise en utilisant la dimension 1 avec la dimension 2 (voir figure 4) ou 3 (Annexe 6).

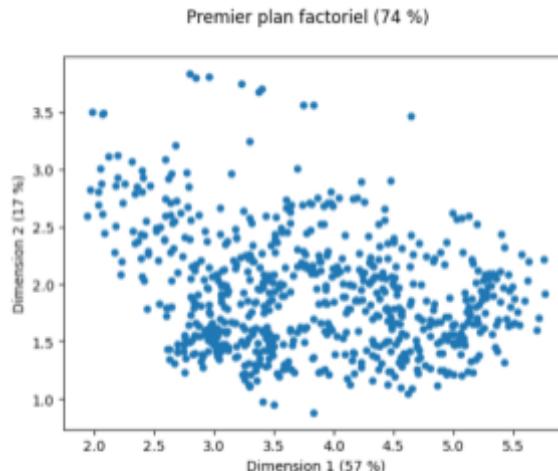


Figure 4 - Nuages de points avec les dimensions 1 et 2

Cette étape avec l'ACP est importante pour l'exploitation de la base de données, en utilisant les K-Means. Pour comprendre mieux ce que représentent les dimensions 1 et 2, nous avons tracé ce que nous appelons un “ cercle de corrélation ”. (voir annexe 6)

4.2-La méthode K-Means (*K-Moyennes*)

L'algorithme K-Means permet de regrouper en un nombre K de clusters (regroupements) distincts les données (voir annexe 2). Dans notre étude, l'utilisation de l'ACP était la première étape du décryptage de la base de données. Cet algorithme nous a permis de nous faire une idée des éléments sur lesquels nous pouvons nous appuyer afin d'extraire les résultats que nous attendons.

Pour réaliser la méthode des K-Means, il faut initialiser des points en utilisant la méthode “K-Means++”, Elle vise à choisir les points de départ initiaux de manière à améliorer la convergence de l'algorithme K-Means et à obtenir des clusters de meilleure qualité.

Elle se divise en plusieurs étapes, tout d'abord il faut sélectionner un premier centroïde de manière aléatoire dans les points disponibles. Ensuite calculer la distance entre chaque point et le centroïde de plus proche déjà sélectionné. Cette distance mesure la "qualité" d'un point en tant que candidat pour le prochain centroïde.

Pour la prochaine étape, il faut sélectionner le prochain centroïde de manière pondérée, en choisissant un point de données avec une probabilité proportionnelle à la distance au centroïde le plus proche. Cela signifie que plus un point est éloigné du centroïde le plus proche, plus il a de chances d'être sélectionné comme prochain centroïde.

Il nous faut répéter les étapes 2 et 3 jusqu'à ce que tous les centroïdes soient sélectionnés. Après avoir fait cela, nous possédonns les centroïdes initiaux pour l'algorithme K-Means. Cela nous permet ensuite de passer à l'étape de regroupement (clustering), où nous attribuons chaque point de données au centroïde le plus proche et itérer jusqu'à ce que la convergence soit atteinte.

Cette méthode aide à répartir les centroïdes de manière plus équilibrée dans l'espace des données, ce qui peut améliorer la convergence de l'algorithme K-Means et éviter les résultats sous-optimaux. Cela permet également de réduire la sensibilité de l'algorithme aux conditions initiales.

Après utilisation, on trace une courbe qui va nous servir d'application de la méthode du "coude". Sur cette courbe, on recherche le point le plus anguleux entre deux intervalles, permettant de déterminer le nombre de clusters optimal que nous produirons avec la méthode K-Means.

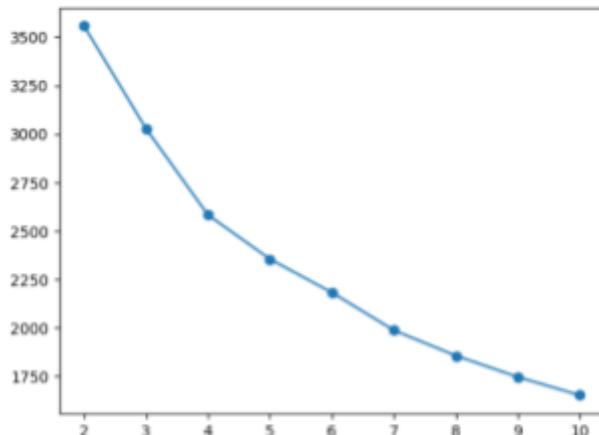


Figure 5 - Méthode du coude (elbow)

Ici, on voit que le "coude" est situé sur l'abscisse 4, on choisira donc de faire 4 clusters.

L'étape suivante est d'utiliser l'algorithme K-Means: il va parcourir l'ensemble des données et affecter à chaque élément des 3 dimensions (que nous avons retenues en utilisant l'ACP) un cluster étiqueté par un nombre entre 0 et 3.

	Dim1	Dim2	Dim3	labels_kmeans
0	3.912831	2.665004	-0.372297	0
1	3.952900	2.263444	-0.774588	0
2	4.383539	1.874272	-0.489254	0
3	2.933331	1.573114	-0.204273	3
4	2.743086	1.476750	-0.297398	3
...
681	2.753948	1.480291	-0.432457	3
682	2.799314	1.493896	-1.180147	3
683	3.258882	1.292053	-0.621750	3
684	3.119590	1.453788	-0.677135	3
685	5.153457	2.216067	-0.630034	2

Figure 6 - Tableau des dimensions de l'ACP et des nombres de regroupement

Une fois les clusters affectés à notre tableau des dimensions, on peut afficher, dans les nuages de points créées plus haut, la répartition de ces derniers, chaque cluster étant représenté par une couleur qui lui est propre. (voir figures ci-après)

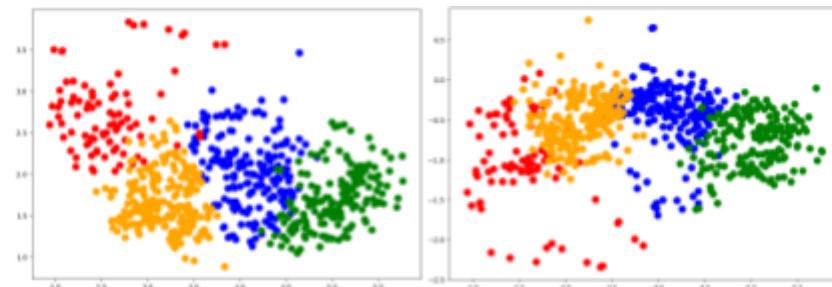


Figure 7 - Graphique représentant les regroupement sur Dim1 et Dim2

Figure 8 - Graphique représentant les regroupement sur Dim1 et Dim3

Ensuite, nous pouvons afficher ces clusters par rapport à la quantité pour chaque pigment. (autre affichage disponible, voir annexe 8)

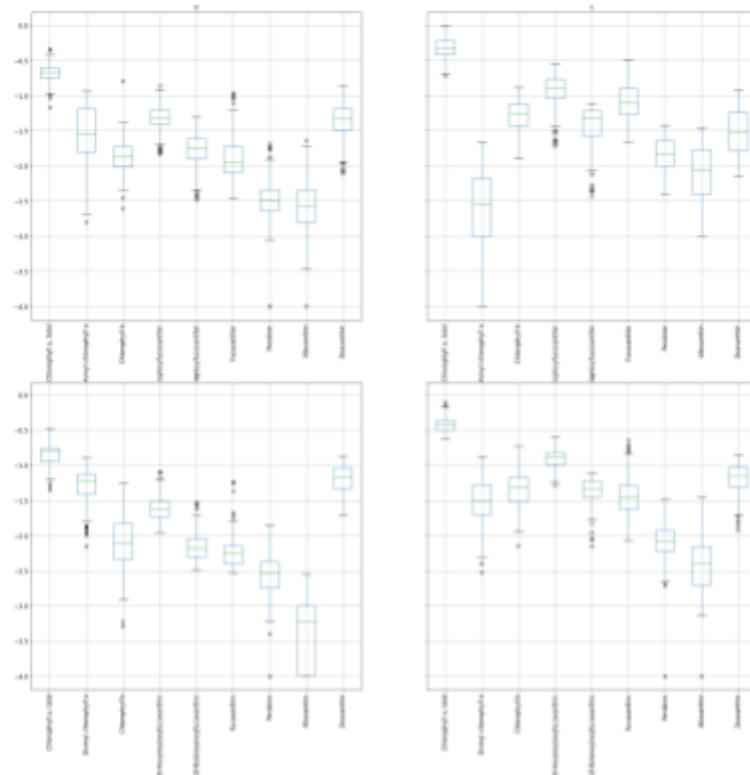


Figure 9 - Représentation des clusters par rapport à leurs quantités de pigments

C'est maintenant que ces graphiques et l'entièreté de notre travail prennent leur sens. En effet, nous sommes désormais capables d'afficher sur une carte du monde la répartition des différents pigments.

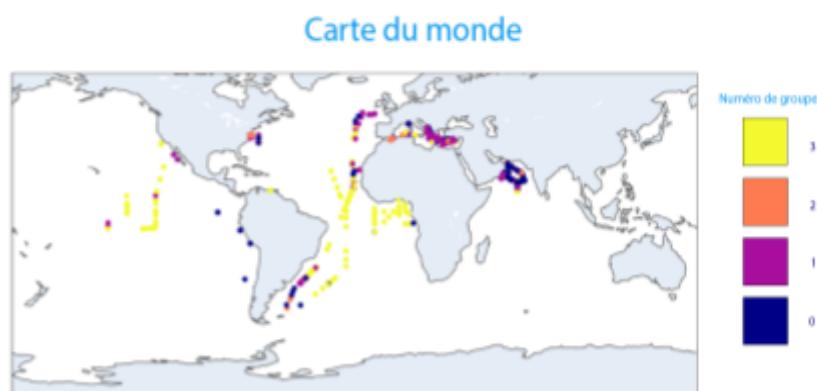


Figure 10 - Répartition des regroupements sur la carte du monde

5- Interprétation des résultats

Après les différentes méthodes de classifications opérées sur la base donnée, nous pouvons observer des résultats très satisfaisants. Nous avons pu constater que la classe 3 représente les Cyanobactéries car cette classe se trouve être riche en divinyle chlorophylle A et en Zeaxanthin mais à l'inverse elle se trouve être très pauvre en chlorophylle A. De son côté, la classe 0 est riche en chlorophylle a et en fucoxanthin, ce qui représente les diatomées. Puis les classes 1 et 2 sont des regroupements mixtes de phytoplanctons, la 2 représente des phytoplanctons de taille moyenne car elle est moyennement riche en chlorophylle à et la classe 1 quant à elle représente des phytoplanctons de petites tailles.

Mais nous avons aussi obtenu des résultats sur la répartition géographiques de ces espèces.

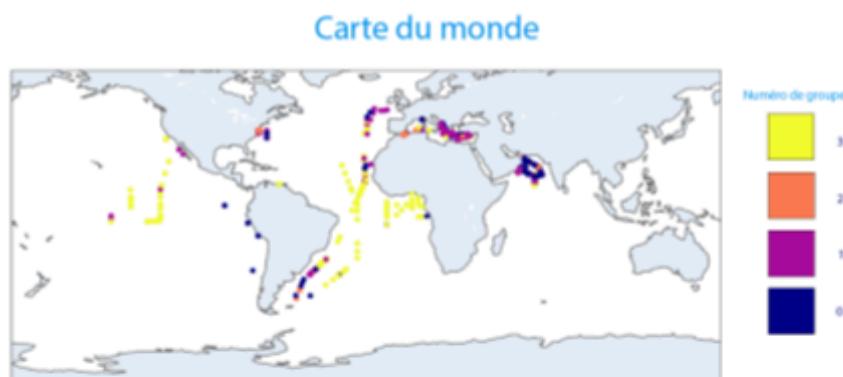


Figure 11 - Répartition des groupes dans le monde

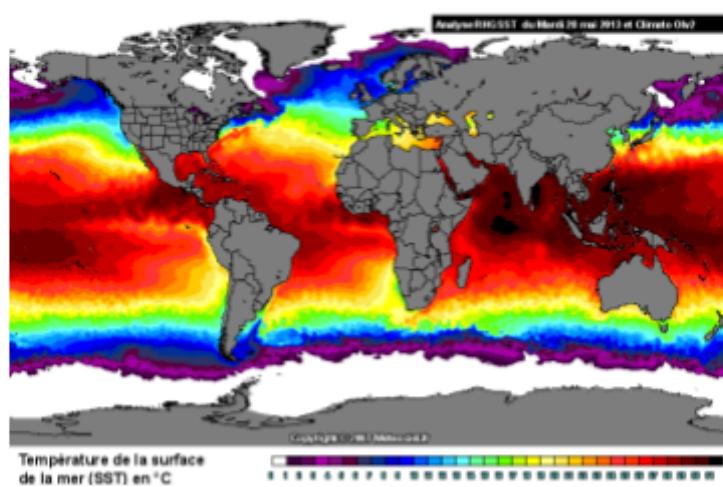


Figure 12 - Carte du monde représentant les températures des océans

Par exemple, nous constatons que la classe 3 représentant les diatomées est présente dans toutes les eaux, qu'elles soient chaudes ou froides. Ce qui correspond bien à ce que nous avions trouvé lors de nos recherches préalables sur le sujet.

6-Appports du projet

Ce projet a été une expérience instructive à la fois sur le plan personnel et technique.

6.1-Appports personnels

Sur le plan personnel, il nous a permis d'approfondir nos connaissances en matière de machine learning et d'explorer les possibilités offertes par l'application de ces méthodes à l'étude des phytoplanctons. Nous avons pu développer nos compétences en manipulation de données, en pré-traitement et en interprétation des résultats. Cette expérience nous a également sensibilisé à l'importance de la recherche marine et de la biodiversité des écosystèmes marins.

6.2-Apports techniques

D'un point de vue technique et pratique, nous avons pu mettre en pratique différentes méthodes et algorithmes avancés. L'utilisation de l'algorithme de l'ACP pour réduire la dimensionnalité des données a été particulièrement pertinente pour explorer la structure sous-jacente des données et identifier les principales composantes qui expliquent la variance. De plus, l'application de l'algorithme des K-Means a facilité la classification des phytoplanctons en regroupements distincts, permettant ainsi une analyse comparative approfondie.

Le pré-traitement des données a notamment joué un rôle crucial dans la qualité et l'interprétation des résultats. En éliminant les données manquantes, en supprimant les variables non pertinentes et en traitant les valeurs aberrantes, nous avons pu obtenir une base de données plus fiable et cohérente. Ces étapes de préparation des données ont été essentielles pour garantir la validité de nos analyses ultérieures.

7- Conclusion

En guise de conclusion, ce projet présente des résultats prometteurs pour la surveillance environnementale et la compréhension des phytoplanctons dans les écosystèmes marins. De plus, l'utilisation du Machine Learning pour l'étude de ces organismes est une initiative innovante qui permet de mieux comprendre la dynamique de ces organismes dans les océans et leur rôle crucial dans l'écosystème marin. Ainsi, l'algorithme des K-means couplé à l'analyse des données satellitaires de couleur de l'océan offre une approche efficace et rapide pour analyser de grandes quantités de données complexes et extraire des informations significatives sur les phytoplanctons et leurs pigments. De plus, il est possible d'avoir accès au code python qui a été utilisé afin d'obtenir ces résultats.

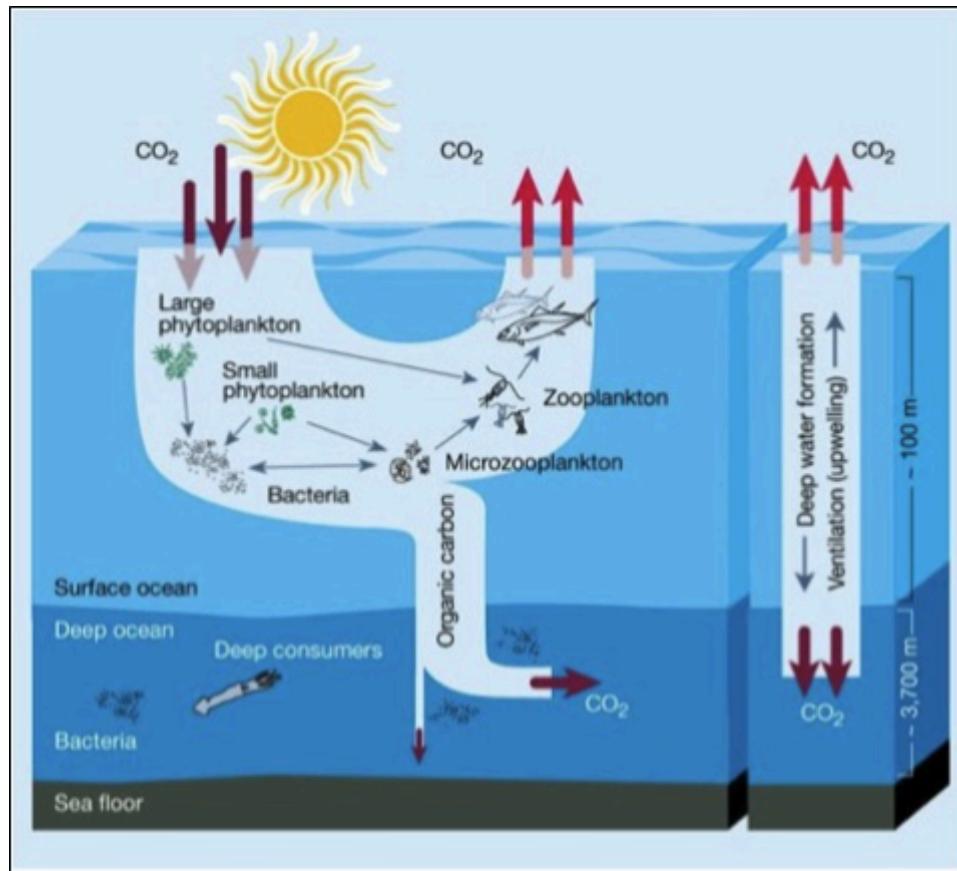
Remerciements

Tout d'abord nous tenons à remercier notre école qui nous donne la possibilité, en classe préparatoire, de pouvoir déjà faire des projets qui nous tiennent à cœur et qui nous immangent dans notre future spécialité.

Ensuite nous souhaitons aussi remercier M. Roucou notre Directeur d'Etudes pour son implication dans le suivi de nos enseignements lors de cette année. Il a montré un grand sens de l'écoute et de l'accompagnement.

Et enfin mais surtout nous voulions absolument remercier notre professeur référent M. El Hourany qui a proposé et supervisé ce projet. Il a été très pédagogue, nous expliquant les démarches à suivre pour le bon déroulement de ce projet et a pris son temps pour nous aider. Mais surtout il nous a proposé un projet intéressant dans l'air du temps qui nous a permis d'en apprendre plus sur le machine learning, l'informatique, mais aussi sur le phytoplancton et notre planète en général.

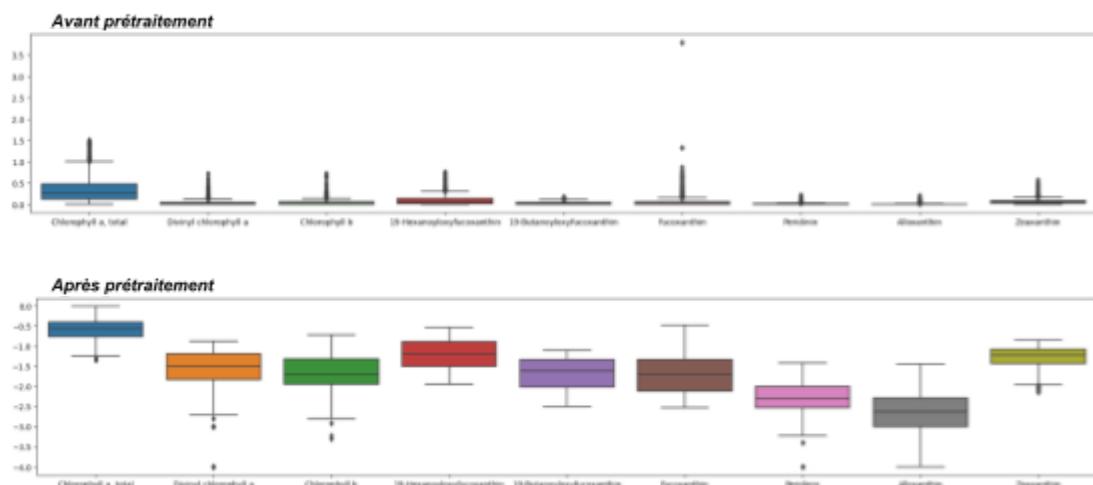
Annexes:



Annexe 1: Schéma des interactions biologiques marines

	Chlorophyll a, total	Dinvinyl chlorophyll a	Chlorophyll b	18-Hexanoyloxyfucoxanthin	18-Butanoyloxyfucoxanthin	Fucoxanthin	Peridinin	Alloxanthin	Zeaxanthin
11	-0.756962	-1.924553	-1.743739	-1.375372	-1.742090	-1.782146	-2.450306	-2.396349	-1.940468
14	-0.723010	-1.496555	-1.800571	-1.479058	-2.058534	-1.547433	-1.906438	-2.533134	-1.636533
16	-0.694369	-1.117057	-1.881382	-1.556185	-2.151720	-1.723676	-2.396186	-2.627770	-1.330647
84	-0.369562	-0.924006	-1.088385	-1.142882	-1.485132	-1.338621	-1.765508	-1.673368	-1.236010
85	-0.297863	-0.914096	-1.089063	-1.036932	-1.449284	-1.226689	-1.587304	-1.634644	-1.129636
...
9136	-0.318799	-1.408895	-1.026872	-0.806875	-1.124939	-1.214670	-2.000000	-2.301030	-1.076721
9144	-0.300162	-1.968607	-1.107905	-0.767004	-1.229148	-1.552842	-1.508638	-2.688970	-0.869666
9146	-0.463442	-1.327902	-1.070581	-0.841638	-1.229148	-1.568636	-2.045757	-3.000000	-1.113509
9147	-0.340084	-1.443697	-1.327902	-0.737549	-1.244125	-1.283997	-2.000000	-3.000000	-1.167491
9477	-1.180104	-1.666653	-1.951947	-1.876475	-2.115771	-2.401209	-2.809668	-3.091515	-1.465466

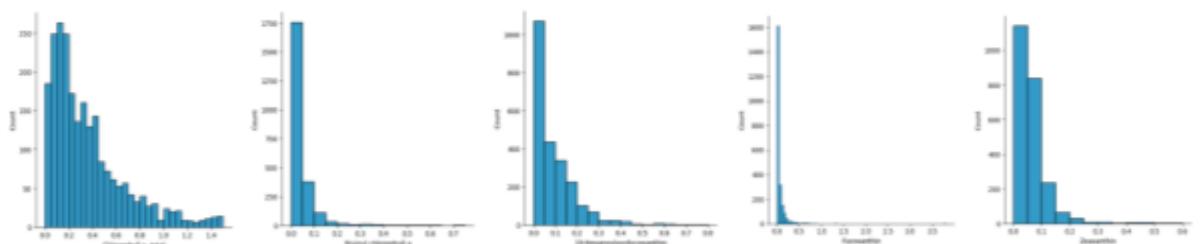
Annexe 2: Base de données triée



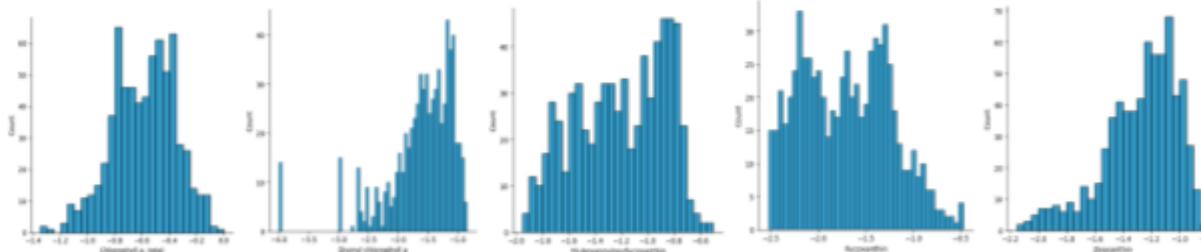
Annexe 3: Répartition des valeurs de concentration des pigments avant et après prétraitement

Définition avant et après traitement des données

AVANT

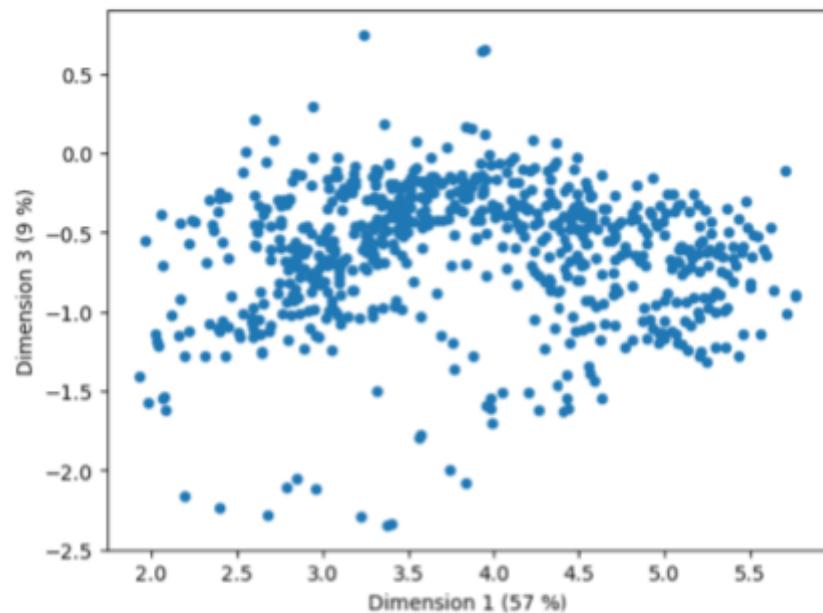


APRES

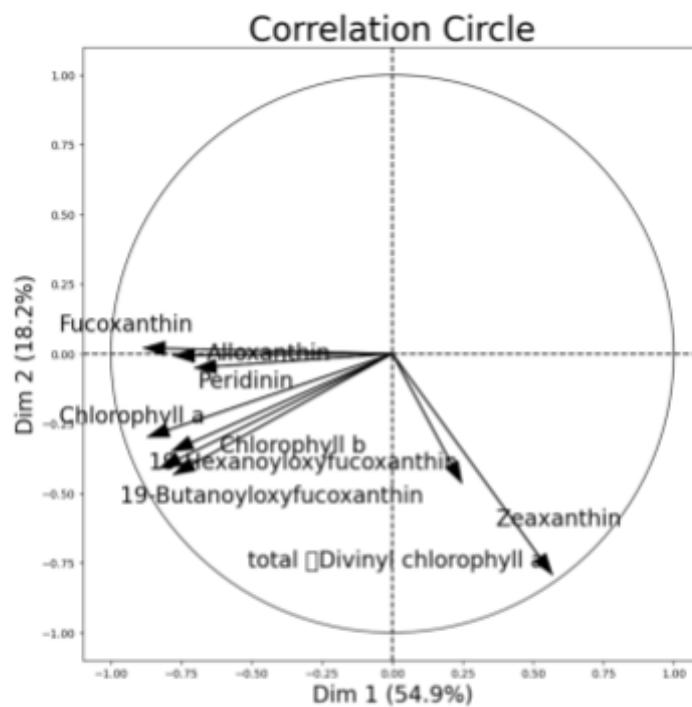


Annexe 4 - Différences entre les valeurs de concentration de pigment avant et après traitement

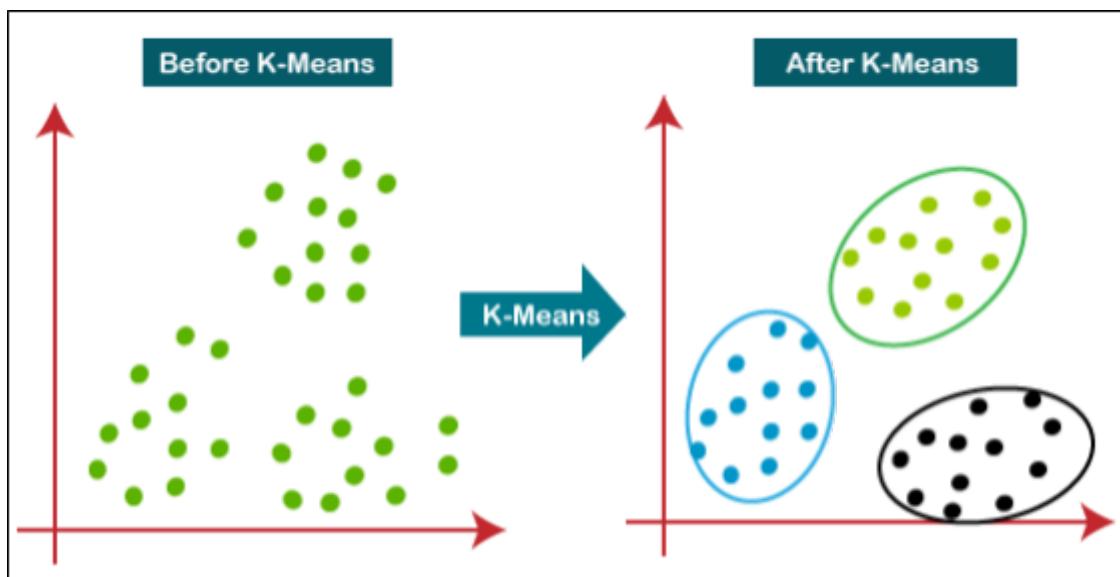
Premier plan factoriel (66 %)



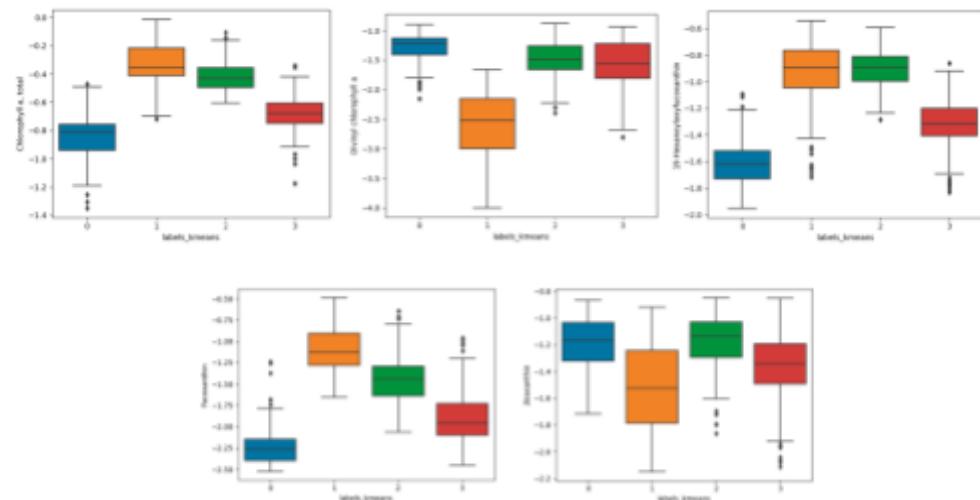
Annexe 5 - Nuages de points avec les dimensions 1 et 3



Annexe 6 - Cercle de corrélation entre Dim 1 et Dim 2



Annexe 7 - Principe de fonctionnement de la méthode "K-MEANS"



Annexe 8 - Représentation des pigments par rapport à leur composition dans chaque cluster